Lead score case study

BY- JAYA KUMARI MODI

Problem Statement

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Business Objective:

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use

Solution Methodology

- Data cleaning and data manipulation.
  1. Check and convert yes/NO categorical variables to 0/1 numerical variable.
  2. Check and handle NA values and missing values.
  3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
  4. Imputation of the values, if necessary.
  5. Check and handle outliers in data.
- EDA
  1. Univariate data analysis: value count, distribution of variable etc.
  2. Bivariate data analysis: pattern between the variables etc.
- Dummy Variables & Feature Scaling and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Dropping features with higher p-value or VIF value
- Re-make the model until above conditions are satisfied.
- Plot ROC curve to get optimal cut-off point.
- Calculate accuracy, sensitivity, specificity of model.
- Validation of the model with test data-set
- Final Model conclusion.

Data Manipulation

At beginning data contained, Total Number of Rows =37, Total Number of Columns =9240.

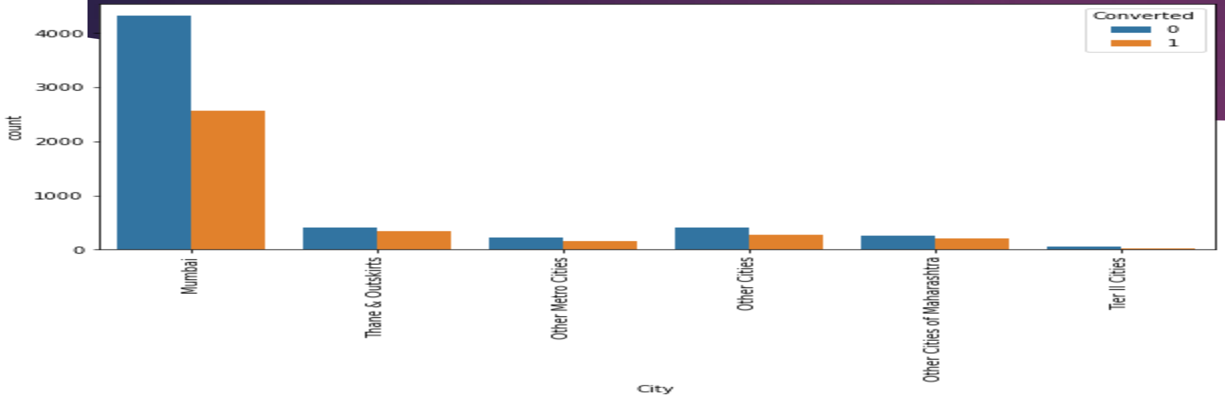
▶ Single value features like "Magazine", "Receive More Updates About Our Courses", "Update me on Supply", Chain Content", "Get updates on DM Content", "I agree to pay the amount through cheque" etc. have been dropped.

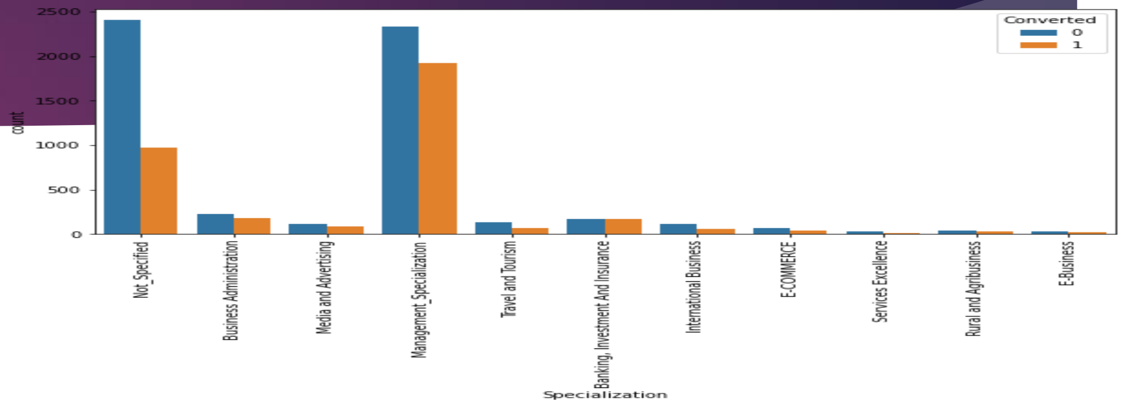
▶ After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: "Do Not Call", "Do not email", "What matters most to you in choosing course", "Search", "Newspaper Article", "X Education Forums", "Newspaper", "Digital Advertisement" etc.

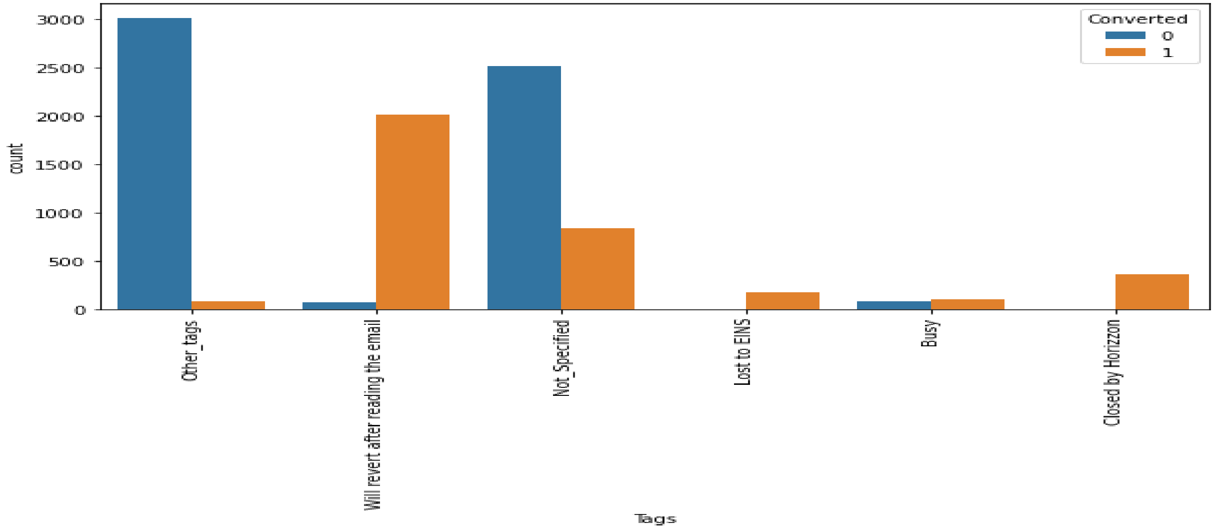
▶ Dropping the columns having more than 45% as missing value such as 'How did you hear about X Education' and 'Lead Profile'.

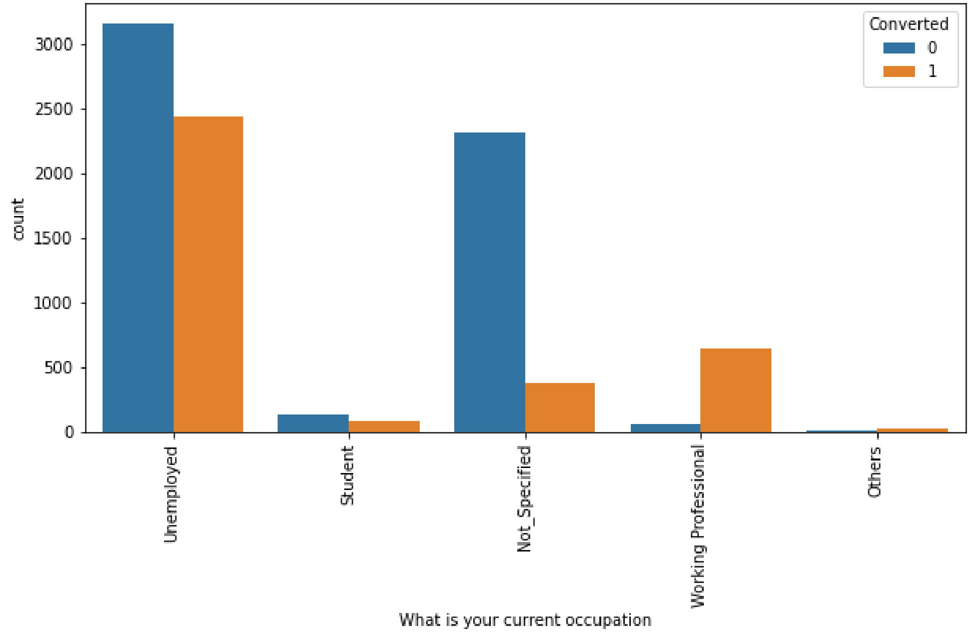
▶Dropping some rows nearly 1.48% where values are missing.

Data Analysis- categorical variables

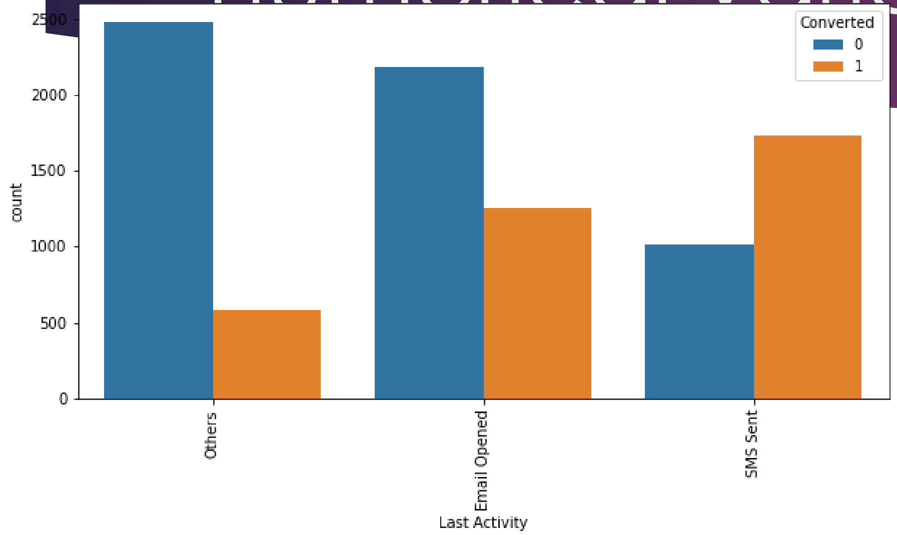


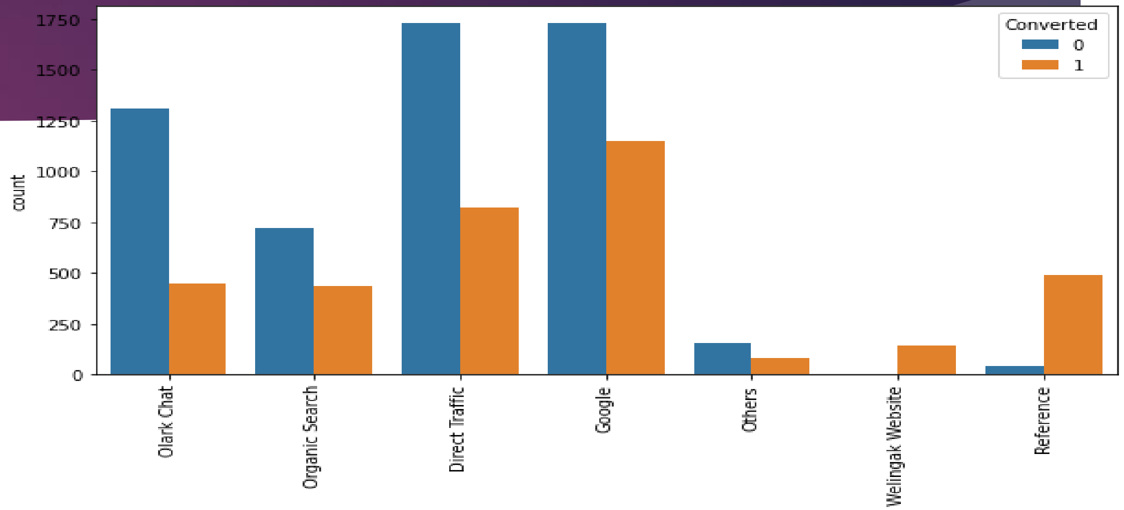


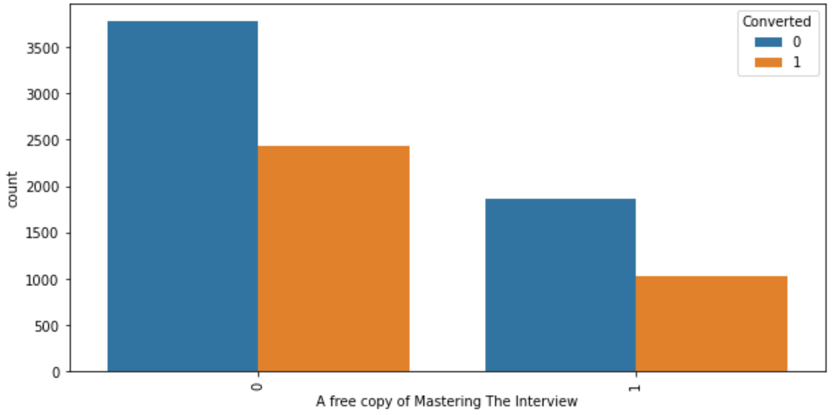




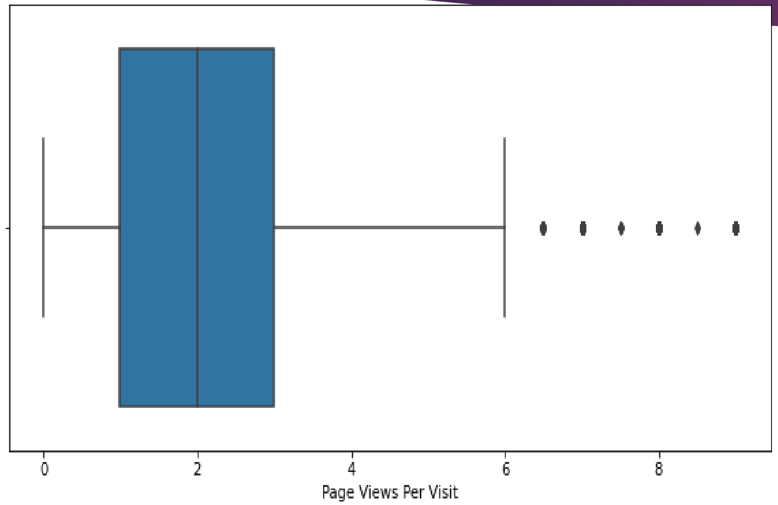
Data Analysis- categorical + discrete numerical variable

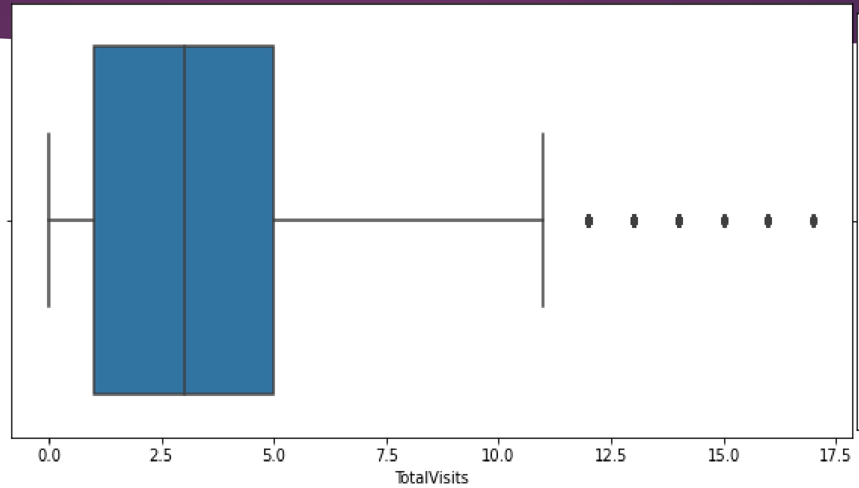


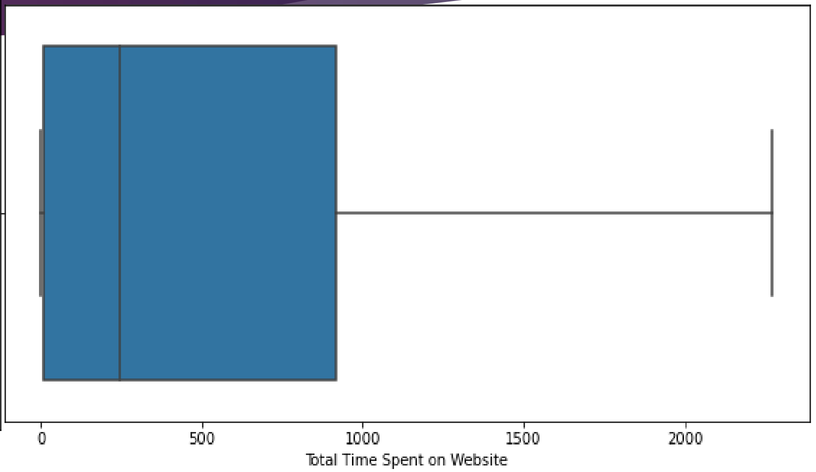


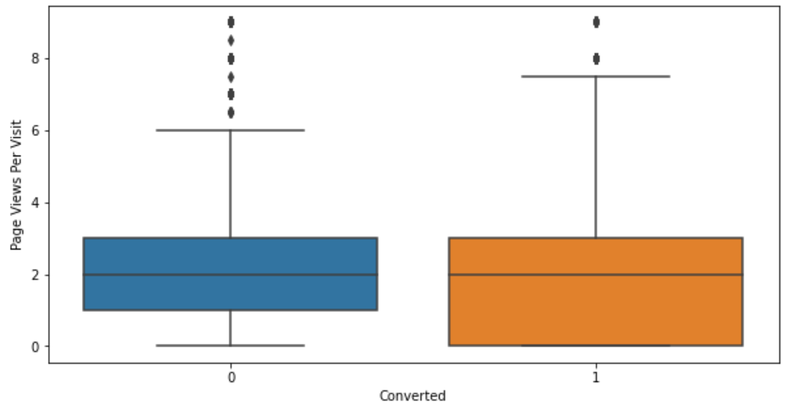


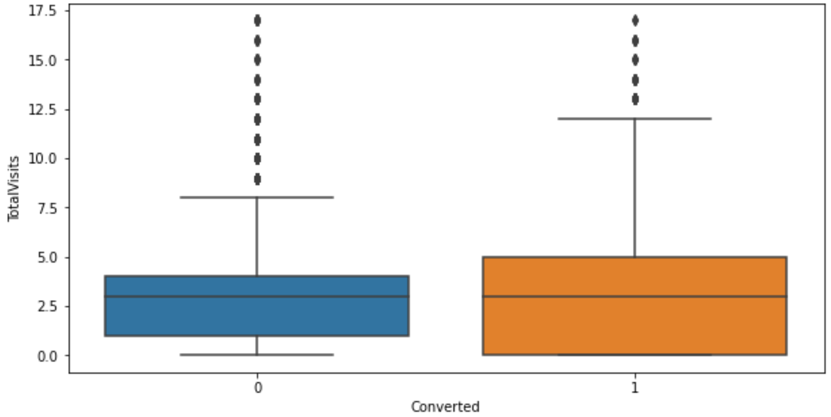
Data Analysis- continuous numerical variables

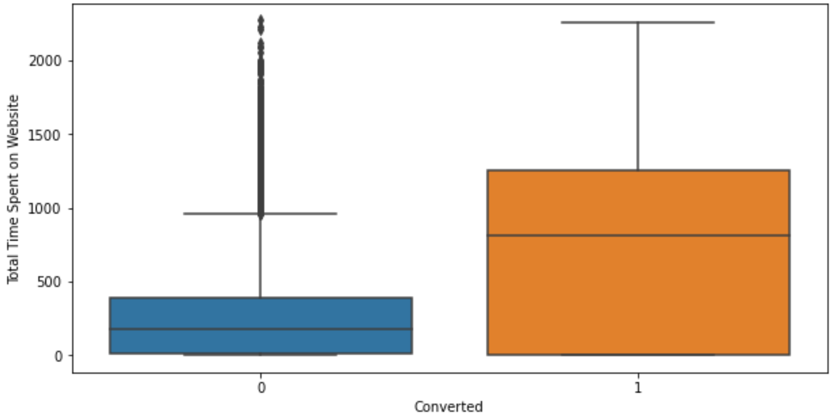












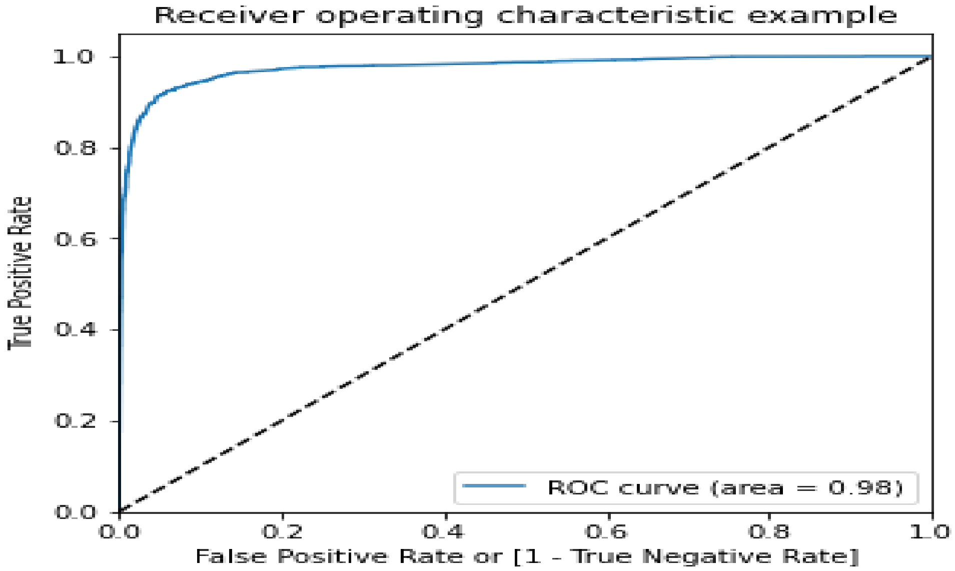
Data Conversion

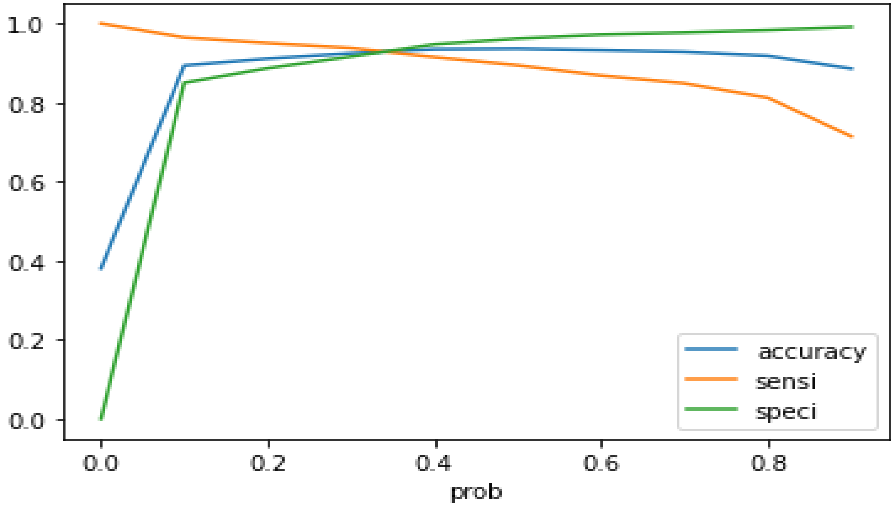
- Numerical Variables are Normalised
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 8953
- Total Columns for Analysis: 46

Model Building

- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5
- Predictions on test data set
- Overall accuracy 93%

ROC curve





Finding Optimal Cut off Point

- Optimal cut off probability is that probability where we get balanced sensitivity and specificity.

- From the second graph it is visible that the optimal cut off is at 0.35.

Conclusion

Features	coef
Tags_Closed by Horizzon	6.7735
Tags_Lost to EINS	5.6049
Tags_Will revert after reading the email	4.4724
Lead Origin_Lead Add Form	3.6828
Last Notable Activity_SMS Sent	3.0727
Total Time Spent on Website	1.0814
Last Notable Activity_Email Opened	1.0366
Lead Source_Olark Chat	0.9617
Tags_Busy	0.7109
Lead Source_Direct Traffic	-0.5416
const	-2.4789
Tags_Other_tags	-2.9731

- It can be concluded that above mentioned features contribute highly for model preparation and also the coefficient is also mentioned along with it.

Conclusion

- Final Observation obtained for Train & Test:

Train Data:

- Accuracy : 93.31%
- Sensitivity :92.41%
- Specificity :93.86%

Test Data:

- Accuracy : 92.85%
- Sensitivity : 90.39%
- Specificity : 94.33%

- This model seems to predict the conversion rate very well and can be used for further analysis by X company.